An evolutionary-based term reduction approach to bilingual clustering of Malay-English corpora

ABSTRACT

The document clustering process groups the unstructured text documents into a predefined set of clusters in order to provide more information to the users. There are many studies conducted in cluster-ing monolingual documents. With the enrichment of current technologies, the study of bilingual clustering would not be a problem. However clustering bilingual document is still facing the same problem faced by a monolingual document clustering which is the "curse of dimensionality". Hence, this encourages the study of term reduction technique in clustering bilingual documents. The objective in this study is to study the effects of reducing terms considered in clustering bilingual corpus in parallel for English and Malay documents. In this study, a genetic algorithm (GA) is used in order to reduce the number of feature selected. A single-point crossover with a crossover rate of 0.8 is used. Not only that,this study also assesses the effects of applying different mutation rate (e.g., 0.1 and 0.01) in selecting the number of features used in clustering bilingual documents. The result shows that the implementation of GAdoes improve the clustering mapping compared to the initial clustering mapping. Not only that, this study also discovers that GA with a mutation rate of 0.01 produces the best parallel clustering mapping results compared to GA with a mutation rate of 0.1.